

# From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction

Simona Cocco<sup>1</sup>, Remi Monasson<sup>2</sup>, Martin Weigt<sup>3,4,\*</sup>

**1** Laboratoire de Physique Statistique de l'Ecole Normale Supérieure - UMR 8550, associé au CNRS et à l'Université Pierre et Marie Curie, 24 rue Lhomond, 75005 Paris, France

**2** Laboratoire de Physique Théorique de l'Ecole Normale Supérieure - UMR 8549, associé au CNRS et à l'Université Pierre et Marie Curie, 24 rue Lhomond, 75005 Paris, France

**3** Université Pierre et Marie Curie, UMR 7238 - Laboratoire de Génomique des Microorganismes, 15 rue de l'Ecole de Médecine, 75006 Paris, France

**4** Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy

\* E-mail: martin.weigt@upmc.fr

## Abstract

Various approaches have explored the covariation of residues in multiple-sequence alignments of homologous proteins to extract functional and structural information. Among those are principal component analysis (PCA), which identifies the most correlated groups of residues, and direct coupling analysis (DCA), a global inference method based on the maximum entropy principle, which aims at predicting residue-residue contacts. In this paper, inspired by the statistical physics of disordered systems, we introduce the Hopfield-Potts model to naturally interpolate between these two approaches. The Hopfield-Potts model allows us to identify relevant 'patterns' of residues from the knowledge of the eigenmodes and eigenvalues of the residue-residue Pearson correlation matrix. We show how the computation of such statistical patterns makes it possible to accurately predict residue-residue contacts with a much smaller number of parameters than DCA. In addition, we show that low-eigenvalue correlation modes, discarded by PCA, are important to recover structural information: the corresponding patterns are highly localized, that is, they are concentrated in few sites, which we find to be in close contact on the three-dimensional protein fold. We also explain why these low-eigenvalue modes, in contrast to the standard principal components, are able to efficiently encode compensatory mutations between pairs of residues.

## Introduction

Thanks to the constant progresses in DNA sequencing techniques, by now close to 4,000 full genomes are sequenced [1], resulting in more than  $2.7 \cdot 10^7$  known protein sequences [2], which are classified into more than 13,000 protein domain families [3], most of them containing in the range of  $10^3 - 10^5$  homologous (i.e. evolutionarily related) amino-acid sequences. These huge numbers are contrasted by only 85,000 experimentally resolved X-ray or NMR structures [4], many of them describing the same proteins. It is therefore tempting to use sequence data alone to extract information about the functional and the structural constraints acting on the evolution of those proteins. Analysis of single-residue conservation offers a first hint about those constraints: Highly conserved positions (easily detectable in multiple sequence alignments corresponding to one protein family) identify residues whose mutations are likely to disrupt the protein function, *e.g.* by the loss of its enzymatic properties. However, not all constraints result in strong single-site conservation. As is well-known, compensatory mutations can happen and preserve the integrity of a protein even if single site mutations have deleterious effects [5,6]. A natural idea is therefore to analyze covariations between residues, that is, whether their variations across sequences are correlated or not. In this context, one introduces a matrix  $\Gamma_{ij}(a,b)$  of residue-residue correlations expressing how much the presence of amino-acid 'a' in position 'i' on the protein is correlated across the sequence data with the presence of another amino-acid, say, 'b', in another position, say 'j'. Extracting information

from this matrix has been the subject of numerous studies over the past two decades, see *e.g.* [5–16].

However, the direct use of correlations for discovering structural constraints such as residue-residue contacts in a protein fold has remained of limited accuracy [5,6,8,11]. More sophisticated approaches to exploit the information included in  $\Gamma$  are based on the *Maximum Entropy* (MaxEnt) [17,18] modeling. The underlying idea is to look for the least constrained statistical model of protein sequences capable of reproducing empirically observed correlations. MaxEnt has been used to analyze many types of biological data, ranging from multi-electrode recording of neural activities [19,20], gene concentrations in genetic networks [21], bird flocking [22] etc. MaxEnt to model covariation in protein sequences was first proposed in a purely theoretical setting by Lapedes *et al.* [7]. It was used (even if not explicitly stated) by Ranganathan and coworkers to generate random protein sequences through Monte Carlo simulations, as a part of an approach called Statistical Coupling Analysis (SCA) [10]. Remarkably, those artificial proteins folded with a substantial probability, which showed that MaxEnt modeling was able to capture structural features essential to the protein family. Recently, one of us proposed, in a series of collaborations, an analytical approach based on the mean-field approximation of statistical physics, called *Direct Coupling Analysis* (DCA), to efficiently compute and exploit this MaxEnt distribution [12,14], related approaches developed partially in parallel are [13,15,16]. Informally speaking, DCA allows for disentangling direct contributions to correlations (resulting from true contacts on the 3D structure) from indirect contributions (mediated through chains of contacts on the protein structure). Hence, DCA offers a much more accurate image of the contact map than  $\Gamma$  itself, and allows to accurately predict protein folds [23–26] and to assemble protein complexes [27,28]. Despite its successes, DCA, and, more generally, MaxEnt modeling raises several concerns. The number of ‘direct coupling’ parameters necessary to define the MaxEnt model over the set of protein sequences, is of the order of  $L^2(q-1)^2$ . Here,  $L$  is the protein length, and  $q = 21$  is the number of amino acids (including the gap). So, for realistic protein lengths of  $L = 50 - 500$ , we end up with  $10^6 - 10^8$  parameters, which have to be inferred from alignments of  $10^3 - 10^5$  proteins. Overfitting the sequence data is therefore a major risk.

Another, and mathematically simpler way to extract information from the correlation matrix  $\Gamma$  is Principal Component Analysis (PCA) [29]. PCA looks for the eigenmodes of  $\Gamma$  associated to the largest eigenvalues. Those modes are the ones contributing most to the covariation in the protein family. Combined with clustering approaches, PCA was applied to the SCA correlation matrix, a variant of the matrix  $\Gamma$  expressing correlations between sites only (and not explicitly the amino-acids they carry) [30,31]. PCA allowed for the identification of groups of correlated (coevolving) residues – termed sectors – each controlling a specific function, in several protein families. A fundamental issue with PCA is the determination of the number of relevant eigenmodes. This is usually done by comparing the spectrum of  $\Gamma$  with a null model, the Marcenko-Pastur (MP) distribution, describing the spectral properties of the sample covariance matrix of a set of independent variables [32]. Eigenvalues larger than the top edge of the MP distribution cannot be explained from sampling noise and are selected, while lower eigenvalues – inside the bulk of the MP spectrum, or even lower – are rejected.

In this article we show that there exists a deep connection between DCA and PCA. To do so we consider the Hopfield-Potts model, an extension of the Hopfield model introduced three decades ago in computational neurosciences [33] to the case of variables taking  $q > 2$  values. The Hopfield-Potts model is based on the concept of patterns, that is, of special directions in the sequence space. Some of those patterns are ‘attractive’, defining ‘ideal’ sequences which real sequences in the protein family try to mimic. In addition, in distinction to the original Hopfield model [33], we introduce ‘repulsive’ patterns, which define regions in the sequence space deprived of real sequences. The statistical mechanics of the inverse Hopfield model, studied in [34] for the  $q = 2$  case and extended here to the generic  $q > 2$  Potts case, shows that it naturally interpolates between PCA and DCA, and allows us to study the statistical issues raised by those approaches exposed above. We show that, in contradistinction with PCA, low eigenvalues and eigenmodes are important to recover structural information about the proteins, and should not be discarded. In addition, we propose refined statistical criteria for the modes to be selected, not based on

the comparison with the MP spectrum. We also study the nature of the eigenmodes (and not only the eigenvalues themselves), and show that they exhibit remarkable features in term of localization: repulsive patterns are strongly localized on (supported by) a few sites, generally found to be in close contact on the three-dimensional structure of the proteins. As for DCA, we show that the dimensionality of the MaxEnt model can be very efficiently reduced with essentially no loss of predictive power for the contact map. These conclusions are established from theoretical arguments, and from the direct application of the Hopfield-Potts model to three sample protein families.

## A short reminder of covariation analysis

Data come in form of a *multiple sequence alignment* (MSA), in which each row gives the amino-acid sequence of one protein, and each column one residue position in these proteins, which is aligned based on amino-acid similarity. Here, the MSA is denoted by  $A = \{a_i^m | i = 1, \dots, L, m = 1, \dots, M\}$  with index  $i$  running over the  $L$  columns of the alignment (residue positions / sites), and  $m$  over the  $M$  sequences, which constitute the rows of the MSA. The amino-acids  $a_i^m$  are assumed to be represented by natural numbers  $1, \dots, q$  with  $q = 21$ , where we include the 20 standard amino acids and the alignment gap '-'.

In our approach, we do not use the data directly, but we summarize them by the amino-acid occupancies in single columns and pairs of columns of the MSA (cf. Methods for data preprocessing),

$$f_i(a) = \frac{1}{M} \sum_{m=1}^M \delta_{a, a_i^m} \quad (1)$$

$$f_{ij}(a, b) = \frac{1}{M} \sum_{m=1}^M \delta_{a, a_i^m} \delta_{b, a_j^m} , \quad (2)$$

with  $i, j = 1, \dots, L$  and  $a, b = 1, \dots, q$ . The Kronecker symbol  $\delta_{a,b}$  equals one for  $a = b$ , and zero else. Since frequencies sum up to one, we can discard one amino-acid value (*e.g.*  $a = q$ ) for each position without losing any information about the sequence statistics. We define the empirical covariance matrix through

$$C_{ij}(a, b) = f_{ij}(a, b) - f_i(a)f_j(b) , \quad (3)$$

with the position index  $i$  running from 1 to  $L$ , and the amino-acid index from 1 to  $q - 1$ . The covariance matrix  $C$  is therefore a square matrix, with  $(q - 1)L$  rows and columns.

## Maximum entropy modeling and direct couplings

The existence of a non-zero covariance between two sites and amino-acids does not necessarily imply that those sites directly interact for functional or structural purposes [8]. The reason is the following [12]: When  $i$  interacts with  $j$ , and  $j$  interacts with  $k$ , also  $i$  and  $k$  will show correlations even if they do not interact. It is thus important to distinguish between *direct* and *indirect* correlations, and to infer *networks of direct couplings*, which generate the empirically observed covariances. This can be done by constructing a (protein-family specific) statistical model  $P(a_1, \dots, a_L)$ , which describes the probability of observing a particular amino-acid sequence  $a_1, \dots, a_L$ . Due to the limited amount of available data, we require this model to reproduce empirical frequency counts for single MSA columns and column pairs,

$$f_i(a_i) = \sum_{\{a_k | k \neq i\}} P(a_1, \dots, a_L) \quad (4)$$

$$f_{ij}(a_i, a_j) = \sum_{\{a_k | k \neq i, j\}} P(a_1, \dots, a_L) , \quad (5)$$

i.e. marginal distributions of  $P(a_1, \dots, a_L)$  are required to coincide with the empirical counts up to the level of position pairs. Beyond this coherence, we aim at the *least constrained* statistical description. The *maximum-entropy principle* [17, 18] stipulates that  $P$  is found by maximizing the entropy

$$S[P] = - \sum_{a_1, \dots, a_L} P(a_1, \dots, a_L) \log P(a_1, \dots, a_L) , \quad (6)$$

subject to the constraints Eqs. (4) and (5). We readily find the analytical form

$$P(a_1, \dots, a_L) = \frac{1}{\mathcal{Z}(\{e_{ij}(a, b), h_i(a)\})} \exp \left\{ \sum_{i < j} e_{ij}(a_i, a_j) + \sum_i h_i(a_i) \right\} , \quad (7)$$

where  $\mathcal{Z}$  is a normalization constant. The MaxEnt model thus takes the form of a (generalized)  $q$ -states Potts model, a celebrated model in statistical physics [35]. The parameters  $e_{ij}(a, b)$  are the direct couplings between MSA columns, and the  $h_i(a)$  represent the local fields (biases) acting on single sites. Their values have to be determined such that Eqs. (4) and (5) are satisfied.

From a computational point of view, however, it is not possible to solve Eqs. (4) and (5) exactly. The reason is that the calculations of  $\mathcal{Z}$  and of the marginals require summations over  $q^L$  microscopic configurations. With  $q = 21$  and typical protein lengths of  $L = 50 - 500$ , the numbers of configurations are enormous, of the order of  $10^{65} - 10^{650}$ . The way out is an approximate determination of the model parameters. The computationally most efficient way found so far is an approximation, called mean field in statistical physics, leading to the approach known as *direct coupling analysis* [14]. Within this mean-field approximation, the values for the direct couplings are simply equal to

$$e_{ij}(a, b) = (C^{-1})_{ij}(a, b) \quad \forall i < j \quad \forall a, b = 1, \dots, q - 1, \quad (8)$$

and  $e_{ij}(a, q) = e_{ij}(q, a) = 0$  for all  $a = 1, \dots, q$ . Note that the couplings can be approximated with this formula in a time of the order of  $L^3(q - 1)^3$ , instead of the exponential time complexity,  $q^L$ , of the exact calculation. On a single desktop PC, this can be achieved in a few seconds to minutes, depending on the length  $L$  of the protein sequences.

The problem can be formulated equivalently in terms of maximum-likelihood (ML) inference. Assuming  $P(a_1, \dots, a_L)$  to be a pairwise model of the form of Eq. (7), we aim at maximizing the log-likelihood

$$\mathcal{L}[\{e_{ij}(a, b), h_i(a)\} | A] = \frac{1}{M} \sum_{m=1}^M \log P(a_1^m, \dots, a_L^m) \quad (9)$$

of the model parameters  $\{e_{ij}(a, b), h_i(a)\}$  given the MSA  $A$ . This maximization implies that Eqs. (4) and (5) hold. In the rest of the paper, we will adopt the point of view of ML inference, cf. the details given in Methods.

Once the direct couplings  $e_{ij}(a, b)$  have been calculated, they can be used to make predictions about the contacts between residues. More details of how these predictions are made can be found in the Methods Section. In [14], it was shown that the predictions for the residue-residue contacts in proteins are very accurate. In other words, DCA allows to find a very good estimate of a partial contact map from sequence data only. Subsequent works have shown that this contact map can be completed by embedding it into three dimensions [23, 24].

### Pearson correlation matrix and principal component analysis

Another way to extract information about groups of correlated residues is the following. From the covariance matrix  $C$  given in Eq. (3), we construct the Pearson correlation matrix  $\Gamma$  through the relationship

$$\Gamma_{ij}(a, b) = \sum_{c, d=1}^{q-1} (D_i)^{-1}(a, c) C_{ij}(c, d) (D_j)^{-1}(d, b) , \quad (10)$$

where the matrices  $D_i$  are the square roots of the single-site correlation matrices, *i.e.*

$$C_{ii}(a, b) = \sum_{c=1}^{q-1} D_i(a, c) D_i(c, b) . \quad (11)$$

This particular form of the Pearson correlation matrix  $\Gamma$  in Eq. (10) results from the fact that we have projected the  $q$ -dimensional space defined by the amino-acids  $a = 1, \dots, q$  onto the subspace spanned by the first  $q - 1$  dimensions. Alternative projections lead to modified but equivalent expressions of the Pearson matrix, cf. the Supporting Information. Informally speaking, the correlation  $\Gamma_{ij}(a, b)$  is a measure of comparison of the empirical covariance  $C_{ij}(a, b)$  with the single-site fluctuations taken independently. Hence,  $\Gamma$  is normalized and coincides with the  $(q-1) \times (q-1)$  identity matrix on each site:  $\Gamma_{ii}(a, b) = \delta_{a,b}$ .

We further introduce the eigenvalues and eigenvectors ( $\mu = 1, \dots, L(q-1)$ )

$$\sum_{j=1}^L \sum_{b=1}^{q-1} \Gamma_{ij}(a, b) v_{jb}^\mu = \lambda_\mu v_{ia}^\mu , \quad (12)$$

where the eigenvalues are ordered in decreasing order  $\lambda_1 > \lambda_2 > \dots > \lambda_{L(q-1)}$ . The eigenvectors are chosen to form an ortho-normal basis,

$$\sum_{ia} v_{ia}^\mu v_{ia}^\nu = L \delta_{\mu,\nu} , \quad (13)$$

for all  $\mu, \nu = 1, \dots, L(q-1)$ . Principal component analysis consists in keeping only the eigenmodes contributing most to the correlations, *i.e.* with the largest eigenvalues, and in discarding all the other eigenvectors. Hence, the directions of maximum covariation of the residues are identified.

PCA is also at the core of principal coordinate analysis or classical scaling [36], which maps the variables considered (here, the pairs  $(i, a)$ ) onto points in a low-dimensional space in such a way that the distance between the points is indicative of the degree of correlation between the attached variables: the closer the points, the more correlated the variables. Such representations are useful to identify clusters of highly correlated variables. Let  $p$  be the number of selected modes. Each variable  $i = 1, 2, \dots, L; a = 1, 2, \dots, q-1$  defines a point in the  $p$ -dimensional space, with coordinates  $\vec{r}_{i,a} = (\sqrt{\lambda_1} v_{ia}^1, \sqrt{\lambda_2} v_{ia}^2, \dots, \sqrt{\lambda_p} v_{ia}^p)$ . When  $p = L(q-1)$ , then all modes are selected and

$$\frac{1}{2L} (\vec{r}_{i,a} - \vec{r}_{j,b})^2 = \frac{1}{2} \Gamma_{ii}(a, a) + \frac{1}{2} \Gamma_{jj}(b, b) - \Gamma_{ij}(a, b) = 1 - \Gamma_{ij}(a, b) , \quad (14)$$

which shows that closest points indeed correspond to largest correlations. When  $p < L(q-1)$ , the left hand side of (14) is the 'best'  $p$ -dimensional approximation to its right hand side.

PCA was used in the context of protein residue covariation by Ranganathan and coworkers [6]. In their approach, called statistical coupling analysis (SCA), a modified covariance matrix,  $\tilde{C}^{SCA}$ , is introduced :

$$\tilde{C}_{ij}^{SCA}(a, b) = w_i^a C_{ij}(a, b) w_j^b \quad (15)$$

where the weights  $w_i^a$  favor positions  $i$  and residues  $a$  with high conservation. Then the amino-acid indices are contracted to define the effective covariance matrix,

$$\tilde{C}_{ij}^{SCA} = \sqrt{\sum_{a,b} \tilde{C}_{ij}^{SCA}(a, b)^2} . \quad (16)$$

The entries of  $\tilde{C}^{SCA}$  depend on the residue positions  $i, j$  only. In a variant of SCA the amino-acid information is directly contracted at the level of the sequence data. A binary variable is associated to each site: it is equal to one in sequences carrying the consensus amino-acid, to zero otherwise [30]. Principal component analysis can then be applied to the  $L$ -dimensional  $\tilde{C}^{SCA}$  matrix, and used to define clusters of correlated sites.

## Results

To bridge these two approaches – DCA and PCA – we introduce the Hopfield-Potts model for maximum likelihood modeling of the sequence distribution, given the residue frequencies  $f_i(a)$  and their pairwise correlations  $f_{ij}(a, b)$ . From a mathematical point of view, the model corresponds to a specific class of Potts models, in which the coupling matrix  $e_{ij}(a, b)$  is of low rank  $p$  compared to  $L(q - 1)$ . It therefore offers a natural way to reduce the number of parameters far below what is required in the mean-field approximation. In addition, the solution of the Hopfield-Potts inverse problem, *i.e.* the determination of the low rank coupling matrix  $e$ , allows us to establish a direct connection with the spectral properties of the correlation matrix  $\Gamma$  and thus with PCA.

Here, we first give an overview over the most important theoretical results for Hopfield-Potts model inference, increasing levels of detail about the algorithm and its derivation are provided in Methods and Supporting Information. Subsequently we discuss the features of the Hopfield-Potts patterns found in three different protein families, and finally assess our capacity to detect residue contacts using sequence information alone.

### Inference with the Hopfield-Potts model

The main idea of this work is to express the matrix  $e_{ij}(a, b)$  in terms of  $p < L(q - 1)$  patterns  $\{\xi_{ia}^\mu, \mu = 1, \dots, p\}$ , and to write

$$e_{ij}(a, b) = \frac{1}{L} \sum_{\mu=1}^p \xi_{ia}^\mu \xi_{jb}^\mu \quad (17)$$

with  $i, j = 1, \dots, L$  being the site indices, and  $a, b = 1, \dots, q$  being amino acids (Potts states); the  $q^{th}$  component of the patterns is set to zero,  $\xi_{i,q}^\mu = 0$ , for compatibility with the mean-field approach exposed above. Note that this matrix, for linearly independent patterns, has rank  $p$ , and it depends only on the  $pL(q - 1)$  parameters  $\xi_{ia}^\mu$ , with  $a \leq q - 1$ , instead of  $\mathcal{O}(L^2(q - 1)^2)$  for the most general case of coupling matrices  $e_{ij}(a, b)$ . It is important to underline that patterns can be

- *real-valued*, and correspond to *attractive* patterns, since sequences  $(a_1, \dots, a_L)$  aligned along these patterns, *i.e.* with large values of  $|\sum_i \xi_{ia_i}^\mu|$ , have increased probabilities under the Hopfield-Potts model, or
- *imaginary-valued*, and lead to a negative prefactor in Eq. (17). These patterns correspond to *repulsive* directions for amino-acid sequences  $(a_i, \dots, a_L)$  [34, 37], since the probability  $P$  now decreases when increasing the alignment score  $|\sum_i \xi_{ia_i}^\mu|$ .

In the following we will allow for mixed models having both attractive and repulsive patterns. As we will see a purely attractive Hopfield-Potts model has a substantially worse performance in predicting residue-residue contacts than such a mixed model.

The pattern values can be computed according to the ML principle, see Methods, and expressed in terms of the eigenvectors of the correlation matrix  $\Gamma$ , which were defined in Eq. (12)

$$\xi_{ia}^\mu = \sqrt{1 - \frac{1}{\lambda_\mu}} \tilde{v}_{ia}^\mu, \quad (18)$$

where

$$\tilde{v}_{ia}^\mu = \sum_{b=1}^{q-1} (D_i)^{-1}(a, b) v_{ib}^\mu. \quad (19)$$

Note that the prefactor  $\sqrt{1 - 1/\lambda_\mu}$  is real for  $\lambda_\mu > 1$ , it vanishes for  $\lambda_\mu = 1$ , and becomes imaginary for  $\lambda_\mu < 1$ . According to the discussion above, large eigenvalues ( $> 1$ ) therefore correspond to attractive

patterns, and small eigenvalues ( $< 1$ ) to repulsive patterns. It is not surprising that  $\lambda = 1$  plays a special role, as it coincides with the mean of the eigenvalues:

$$\frac{1}{L(q-1)} \sum_{\mu} \lambda_{\mu} = \frac{1}{L(q-1)} \sum_{ia} \Gamma_{ii}(a, a) = 1 . \quad (20)$$

Equation (18) defines  $L(q-1)$  different patterns, therefore we now need a rule for selecting the  $p$  'best' patterns. We show in Methods that the contribution of the pattern  $\xi^{\mu}$  to the log-likelihood  $\mathcal{L}$  (9) is a function of the associated eigenvalue only,

$$\Delta\mathcal{L}(\lambda_{\mu}) = \frac{1}{2} \left( \lambda_{\mu} - 1 + \log \lambda_{\mu} \right) . \quad (21)$$

As is shown in Fig. 1, large contributions arrive from both the largest and the smallest eigenvalues, whereas eigenvalues close to unity contribute little. Therefore, we have to select the  $p$  eigenvalues with largest contributions. We define a threshold value  $\theta$  such that there are exactly  $p$  patterns with larger contributions to the log-likelihood:

$$|\{\lambda_{\mu} | \Delta\mathcal{L}(\lambda_{\mu}) > \theta\}| = p ; \quad (22)$$

the  $L(q-1) - p$  patterns with smaller  $\Delta\mathcal{L}$  are omitted in the expression for the coupling, cf. Eq. (17). We look thus for the two positive real roots  $\ell_{\pm}$  ( $\ell_- < 1 < \ell_+$ ) of the equation

$$\Delta\mathcal{L}(\ell_{\pm}) = \theta , \quad (23)$$

and select the  $p_-$  repulsive patterns with  $\lambda_{\mu} < \ell_-$  and the  $p_+$  attractive patterns with  $\lambda_{\mu} > \ell_+$ . The total number of selected patterns is  $p = p_- + p_+$ .

An alternative criterion for pattern selection, built on a Bayesian framework of inference, is proposed in Methods. The criterion consists in estimating the uncertainty on each inferred pattern due to limited sampling (sequences number  $M$  in the MSA), and in selecting patterns with small uncertainties only. Remarkably both criteria are in excellent quantitative agreement in practice, cf. Methods.

## Features of the Hopfield-Potts patterns

We have tested the above inference framework using three protein families, which variable values of protein length  $L$  and sequence number  $M$ :

- The *Kunitz/Bovine pancreatic trypsin inhibitor* domain (PFAM ID PF00014) is a relatively short ( $L = 53$ ) and not very frequent ( $M = 2,143$ ) domain, after reweighting the effective number of diverged sequences is  $M_{eff} = 1,024$  (cf. Eq. (26) in Methods for the definition). Results are compared to the exemplary X-ray crystal structure with PDB ID 5pti [38].
- The bacterial *Response regulator* domain (PF00072) is of medium length ( $L = 112$ ) and very frequent ( $M = 62,074$ ). The effective sequence number is  $M_{eff} = 29,408$ . The PDB structure used for verification has ID 1nxw [39].
- The eukaryotic signaling domain *Ras* (PF00071) is the longest ( $L = 161$ ) and has an intermediate size MSA ( $M = 9,474$ ), leading to  $M_{eff} = 2,717$ . Results are compared to PDB entry 5p21 [40].

To interpret the Hopfield patterns in terms of amino-acid sequences, we first report some empirical observations made for the patterns corresponding to the largest and smallest eigenvalues, *i.e.* to the most likely attractive and repulsive patterns. We concentrate here on one protein family, the Trypsin inhibitor (PF00014). Analogous properties are observed in the other two protein families, as reported in the Supporting Information.

The upper panel of Fig. 2 shows the spectral density. It is characterized by a pronounced peak around eigenvalue 1. The smallest eigenvalue is  $\lambda_m^{PF00014} \sim 0.1$ , the largest is  $\lambda_M^{PF00014} \sim 23$ . Large eigenvalues are isolated from the bulk of the spectrum, small eigenvalues are not.

To characterize the statistical properties of the patterns we define, inspired from localization theory in condensed matter physics, the inverse participation ratio (IPR) of each Hopfield pattern  $\xi^\mu = (\xi_{ia}^\mu)$  through

$$\text{IPR}(\xi^\mu) = \frac{\sum_{i,a} (\xi_{ia}^\mu)^4}{\left(\sum_{i,a} (\xi_{ia}^\mu)^2\right)^2} . \quad (24)$$

IPR values are real and positive for all patterns, be they attractive or repulsive. In addition, IPR range from one for perfectly localized patterns (only one single non-zero component), and  $1/(L(q-1))$  for a completely distributed pattern with uniform entries. IPR is therefore used as a localization measure for the patterns: the inverse,  $1/\text{IPR}(\xi^\mu)$ , is an estimate of the number of pairs  $(i, a)$  on which pattern  $\mu$  has sizable entries  $\xi_{ia}^\mu$ . The lower panel of Fig. 2 shows the presence of strong localization for repulsive patterns (small eigenvalues) and for irrelevant patterns (around eigenvalue 1). A much smaller increase in the IPR is also observed for part of the large eigenvalues.

### Repulsive patterns

In the upper row of Fig. 3 we display the three most localized repulsive patterns (smallest, 3rd and 4th smallest eigenvalues) for the trypsin inhibitor protein (PF00014). All three have two very pronounced peaks and some smaller minor peaks, resulting in IPR values above 0.3. For each of the patterns, the two peaks are of opposite sign, and have highest value for the amino acid cysteine. Actually, for all three vectors, the pairs of peaks identify disulfide bonds, *i.e.* covariant bonds between two cysteines which are, in general, very important for a protein's stability and therefore highly conserved. The fact that the peaks are of opposite sign can be interpreted: the corresponding repulsive patterns forbid amino-acid configurations with a cysteine in one site, but not in the other one, see Discussions. Both residues are co-conserved. Note also that the trypsin inhibitor has only three disulfide bonds, *i.e.* all of them are seen by the most localized repulsive patterns. The second eigenvalues, which has a slightly smaller IPR, is actually found to be a mixture of two of these bonds, *i.e.* it is localized over four positions.

The observation of disulfide bonds is specific to the trypsin inhibitor. In other proteins, also the ones studied in this paper, we find similarly strong localization of the most repulsive patterns, but in different amino acid combinations (Supporting Information). In all these cases, the consequence is a co-conservation of these positions, and they are typically found in direct contact.

### Attractive patterns

The strongest attractive pattern, *i.e.* the one corresponding to the largest eigenvalue  $\lambda^1$ , is shown in the leftmost panel of the lower row of Fig. 3. Its IPR is small ( $\sim 0.003$ ), implying that it is extended over most of the protein. As is shown in the Supporting information, strongest entries in  $\xi_{ia}^1$  correspond to conserved residues and these, even if they are distributed along the primary sequence, tend to form spatially connected and functionally important regions in the folded protein (*e.g.* a binding pocket), cf. left panel of Fig. 4. Clearly this observation is reminiscent of the protein sectors observed in [30], which are found by PCA applied to the before-mentioned modified covariance matrix. Note, however, that sectors are extracted from more than one principal component, and without the use of protein structure.

More characteristic patterns are found for the second and third eigenvalues. As is shown in Fig. 3, they show strong peaks at the extremities of the sequence, which become higher when approaching the



first resp. last sequence position. The peaks are concentrated on the gap symbol. The vectors are actually artifacts of the multiple-sequence alignment: Many sequences start or end with a stretch of gaps, which may have one out of at least three reasons: (1) The protein under consideration does not match the full domain definition of PFAM. (2) The local nature of PFAM alignments has initial and final gaps as algorithmic artifacts, a correction would however render the search tools less efficient. (3) In general, in sequence alignment the extension of an existing gap is less expensive than opening a new gap. The attractive nature of these two patterns, and the equal sign of the peaks, imply that gaps in equilibrium configurations of the Hopfield-Potts model frequently come in stretches, and not as isolated symbols. The finding that there are two patterns with this characteristic can be traced back to the fact that each sequence has two ends, and these behave independently with respect to alignment gaps.

### Theoretical results for localization in the limit case of strong conservation

The main features of the empirically observed spectral and localization properties of Fig. 2 can be found back in the limit case of completely conserved sequences, which is amenable to an exact mathematical treatment. To this end, we consider  $L$  perfectly conserved sites, *i.e.* a MSA made from the repetition of a unique sequence. As is shown in the Supporting Information, the corresponding Pearson correlation matrix  $\Gamma$  has only three different eigenvalues:

- a large and non-degenerate eigenvalue,  $\lambda_+$ , which is a function of  $q$  and  $L$  (and of the pseudocount used to treat the data, see Methods), whose corresponding eigenvector is extended;
- a small and  $(L-1)$ -fold degenerate eigenvalue,  $\lambda_- = (L-\lambda_+)/ (L-1)$ . The corresponding eigenspace is spanned by vectors which are perfectly localized in pairs of sites, with components of opposite signs;
- the eigenvalue  $\lambda = 1$ , which is  $L(q-2)$ -fold degenerate. The eigenspace is spanned by vectors, which are localized over single sites.

For a realistic MSA, *i.e.* without perfect conservation, degeneracies will disappear, but the features found above remain qualitatively correct. In particular, we find in real data a pronounced peak of eigenvalues around 1, corresponding to localized eigenmodes (Fig. 2). In addition, low-eigenvalue modes are found to be strongly localized, and the order of magnitude of  $\lambda_- \simeq 0.09$  is in good agreement with the smallest eigenvalues,  $\simeq 0.1$ , reported for the three analyzed domain families. Finally, the largest eigenmodes are largely extended, as found in the limit case above. Note that the eigenvalues found in the protein spectra, *e.g.*  $\lambda_1 \simeq 23$  for PF00014, are however smaller than in the limit case,  $\lambda_+ \simeq 48$ , due to only partial conservation in the real MSA.

### Residue-residue contact prediction with the Hopfield-Potts model

The most important feature of DCA is its ability to predict pairs of residues, which are distantly positioned in the sequence, but which form native contacts in the protein's tertiary structure, cf. the right panel of Fig. 4. Here, our contact prediction is based on the sampling-corrected Frobenius norm of the  $(q-1)$ -dimensional statistical coupling matrices  $e_{ij}$ , cf. Methods, which in [41] has been shown to outperform the direct-information measure used in [12]. This measure assigns a single scalar value for the strength of the direct coupling between two residue positions. The lower panels of Fig. 5 show, for various values of the number  $p$  of patterns, the performance in terms of contact predictions, where two residues are considered to be in contact if their distance is smaller than 8 Å in the before mentioned exemplary protein crystal structures. The plots show the fraction of true-positives (TP), *i.e.* of native contacts, in between the  $x$  pairs of highest DI, as a function of  $x$  [14]. To include only non-trivial predictions, we require also a minimum separation  $|i-j| > 4$  of at least 5 residues along the protein sequence.

The three upper panels in Fig. 5 show the ratio between the selected pattern contributions to the log-likelihood,  $\sum_{\{\mu|\lambda_\mu \notin (\ell_-, \ell_+)\}} \Delta\mathcal{L}(\lambda_\mu)$ , and its maximal value obtained by including all  $L(q-1)$  patterns,  $\sum_{\mu=1}^{L(q-1)} \Delta\mathcal{L}(\lambda_\mu)$ . A large fraction of patterns can be omitted without any substantial loss in log-likelihood, but with a substantially smaller number of parameters. It is worth noticing that we do not find any systematic benefit of excluding patterns for the contact prediction, but the predictive power decreases initially only very slowly with decreasing pattern numbers  $p$ . For all three proteins, even with  $\sim 128$  patterns, very good contact predictions can be achieved, as compared to 1060-3220 patterns for the full mean-field inference. Almost perfect performance is reached, when the contribution of selected patterns to the log-likelihood is only at 60 – 80% of its maximal value. This could be expected from the fact that patterns corresponding to eigenvalues close to unity are very small in norm, see Eq. 18, and hardly contribute to the couplings.

The discussion of the localization properties of repulsive patterns is corroborated by the results reported in Fig. 6. It compares the performance of the Hopfield-Potts model to predict residue-residue contacts, for the three cases where patterns are selected either according to the maximum entropic contribution criterion, or where only the strongest attractive (largest  $\lambda$ ) or only the strongest repulsive (smallest  $\lambda$ ) patterns are taken into account. It becomes evident that the more accurate contact information is given by the repulsive patterns, it is strongly reduced when considering only attractive patterns, *i.e.* in the case corresponding most closely to PCA. This finding illustrates one of the most significant differences between DCA and PCA: Contact information is provided by the strongly localized eigenvectors of the Pearson correlation matrix  $\Gamma$  in the lower tail of the spectrum.

As discussed in the previous paragraph, patterns with the largest contribution to the log-likelihood are dominated by (and localized in) conserved sites. Attractive patterns favor these sites to jointly assume their conserved values, whereas repulsive patterns avoid configurations where, in pairs of co-conserved sites, only one variable assumes its conserved value, but not the other one. However, we have also seen that an accurate contact prediction requires at least  $\sim 100$  patterns, *i.e.* it goes well beyond the patterns given by strongly conserved sites. In Fig. 4 we show, for the exemplary case of the Trypsine inhibitor, both the 15 sites of highest entry in the most attractive pattern  $\xi^1$  (corresponding to conserved sites), and the first 50 predicted intra-protein contacts using the full mean-field DCA scheme (results for  $p = 512$  are almost identical). It appears that many of the correctly predicted contacts are not included in the set of the most conserved sites. From a mathematical point of view, this is understandable - only variable sites may show strong covariation. From a biological point of view, this is very interesting, since it shows that highly variable residue in proteins are not necessarily functionally unimportant in a protein family, but they may undergo strong co-evolution with other sites, and thus be very important for the structural stability of the protein.

A last remark is necessary concerning the right panel of Fig. 4: Whereas conserved sites (which carry also the largest entries of the pattern with maximum eigenvalue) are collected in one or two spatially connected regions in the studied proteins, this is not necessarily true for all proteins. In particular complex domains with multiple functions and/or multiple conformations may show much more involved patterns. It is, however, beyond the scope of this paper to shed light onto the details of the biological interpretation of the principal components of  $\Gamma$ .

## Discussion

In this paper we have proposed a method to analyze the correlation matrix of residues substitutions across multiple-sequence alignments of homologous proteins, based on the inverse Hopfield-Potts model. Our approach offers a natural interpolation between the spectral analysis of the correlation matrix, carried out in principal component analysis, and maximum entropy approaches which aim at reproducing those correlations within a global statistical model. The inverse Hopfield-Potts model requires to infer “directions” of particular importance in the sequence space, called patterns: The distribution of sequences

belonging to a protein family tends to accumulate along attractive patterns (related to eigenmodes of the correlation matrix with large eigenvalues) and to get depleted along repulsive patterns (related to the low-eigenvalue modes). Contrary to principal component analysis, which discards low-eigenvalue modes, we have shown that repulsive patterns are important to characterize the sequence distribution, and in particular to detect structural properties (contact map) of proteins from sequence data. In addition, we have shown how to infer not only the values of the patterns but also their statistical relevance from the sequence data. To do so we have proposed two criteria, based on maximum-likelihood and Bayesian inference (see Methods), which differ from the usual comparison to the Marcenko-Pastur spectrum. Those criteria and the results of the application of the inverse Hopfield-Potts model to real sequence data confirm that most eigenmodes (with eigenvalues close to unity) can be discarded without affecting considerably the contact prediction. This makes our approach much less parameter-intensive than the full direct coupling approach. We have found empirically that it is sufficient to take into account the patterns contributing to  $\sim 60 - 80\%$  of the log-likelihood to achieve a very contact map prediction.

We have also studied the position-specific nature of patterns, taking inspiration from localization theory in condensed matter physics and random matrix theory (Fig. 3 and Supporting Information, Fig. 6 & 10). Briefly speaking, a pattern is said to be localized if it is concentrated on a few sites of the sequence, and extended (over the sequence) otherwise. We have found that the principal attractive pattern (corresponding to the largest eigenvalue) is extended, with entries of largest absolute value in the most conserved sites (Supporting Information, Fig. 3, 4, 7 & 11). Other strongly attractive patterns can be explained from the presence of extended gaps in the alignment, mostly found at the beginning or at the end of sequences. The other patterns of large likelihood contributions are repulsive, *i.e.* they correspond to small eigenvalues, usually discarded by principal component analysis. Interestingly, these patterns appear to be strongly localized, that is, strongly concentrated in very few positions, which despite their separation along the sequence are found in close contact in the 3D protein structure. To give an example, in the Trypsin inhibitor protein, they are localized in position pairs carrying Cysteine, and being linked by disulfide bonds. Other amino-acid combinations were also found in the other protein families studied here, see Supporting Information. Taking into account only a number  $p$  of such repulsive patterns results in a predicted contact map of comparable quality to the one using maximum-likelihood selection, whereas the same number  $p$  of attractive patterns performs substantially worse (Fig. 6 and Supporting Information, Fig. 5 & 9). The dimensional reduction of the Hopfield-Potts model compared to the Potts model (used in standard DCA) is thus even more increased as many relevant patterns are localized and contain only a few (substantially) non-zero components.

A general finding, supported by a theoretical analysis in the Results section, is that the more repulsive are the patterns, the stronger they are localized, and the more conserved are the residues supporting them. As the number of patterns to be included to reach an accurate contact map is a few hundreds for the protein families considered here, the largest components of the weakly repulsive patterns, *i.e.* with the eigenvalues smaller than, but close to the threshold  $\theta$ , correspond to weakly conserved residues. In consequence many predicted contacts connect low-conservation residues. This statement is apparent from Fig. 4 and Supporting Information, Fig. 8 & 12, which compare the sets conserved sites and the pairs of residues predicted to be in contact by our analysis.

Why are repulsive patterns so successful in identifying contacts, in difference to attractive patterns? To answer this question consider the simple case of a pattern localized in two residues only, say amino-acids  $a$  in position  $i$  and  $b$  in position  $j$ . We further assume that the two non-zero components  $\xi_{ia}$  and  $\xi_{jb}$  have the same amplitude and differ only by sign, *i.e.*  $\xi_{ia} = -\xi_{jb}$ . Now we consider a sequence of amino-acids and ask whether it will be 'aligned', *i.e.* will have a strong projection along the pattern. The

outcome is given in the third column of the following table:

$a_i = a$ ?	$a_j = b$ ?	$ \sum_k \xi_{ka_k} / \xi_{i,a} $	Favored by attractive pattern?	Favored by repulsive pattern?
NO	NO	0	NO	YES
YES	NO	1	YES	NO
NO	YES	1	YES	NO
YES	YES	0	NO	YES

The answer therefore corresponds to a **XOR** (exclusive or) between the presence of the two amino-acids  $a$  and  $b$  on their respective positions  $i$  and  $j$  in the sequence. If the pattern were attractive (cf. fourth column), it would favor sequences where exactly one of the two specified amino-acids is present. For a repulsive pattern (cf. fifth column), unaligned sequences are favored, *i.e.* either both  $a$  and  $b$  are present in positions  $i$  and  $j$ , or none of the two.

In case we assumed equal sign components, *i.e.*  $\xi_{ia} = \xi_{jb}$ , we would have found the following table:

$a_i = a$ ?	$a_j = b$ ?	$ \sum_k \xi_{ka_k} / \xi_{i,a} $	Favored by attractive pattern?	Favored by repulsive pattern?
NO	NO	0	NO	YES
YES	NO	1	NO	YES
NO	YES	1	NO	YES
YES	YES	2	YES	NO

This choice is poor in terms of enforcing covariation in the sequence: Since the couplings (17) are quadratic in the alignment score, an attractive (resp. repulsive) pattern strongly favors (resp. disfavors) the presence of both amino acids  $a$  and  $b$  in positions  $i$  and  $j$ , but it is overall monotonous in the number of correctly present amino acids.

As a conclusion we find that strong covariation can be efficiently enforced only by a repulsive pattern with opposite components (fifth column in the first table above). The acceptance of the NO,NO configuration is desirable, too: It signals the possibility of compensatory mutations, *i.e.* favorable double mutations changing both  $a$  and  $b$  in positions  $i$  and  $j$  to alternative amino acids; it is easy to generalize the above patterns to patterns having more than one favored amino-acid combination (*e.g.* favored pairs  $(a,b)$  and  $(c,d)$  can be enforced by a repulsive pattern with  $\xi_{ia} = -\xi_{ic} = -\xi_{jb} = \xi_{jd}$ ).

This theoretical argument explains why localized repulsive patterns critically encode for covariation. Remarkably the condition that the few, large components of repulsive patterns should sum up to zero agrees well with Fig. 3 and Supporting Information, Fig. 6 & 10. Finally let us emphasize the importance of the prefactor  $\sqrt{1 - \frac{1}{\lambda}}$  of the pattern, cf. Eq. (18), where  $\lambda$  is the eigenvalue attached to the pattern. While this factor is at most equal to 1 for attractive patterns, it can take arbitrarily large values (in modulus) for repulsive patterns. Hence, repulsive patterns can have large very amplitudes (Fig. 3) and provide large contributions to the couplings (and consequently to our contact prediction).

Some aspects of the approach presented in this paper deserve further studies, and may actually lead to substantial improvements of our ability to detect residue contacts from statistical sequence analysis. First the non-independence of sequences in the alignment, *e.g.* due to phylogenetic correlations, should be taken into a more accurate way than done currently by sequence reweighting. The introduction of a large pseudo-count in the data, much larger than the order of  $\sim 1$  expected from a Bayesian theory should also be elucidated. Last, while the use of the Frobenius norm for the coupling  $e_{ij}(a,b)$  (with the average-product correction, see Methods) has proven to be an efficient criterion for contact prediction, it remains unclear if there exist other estimators of contact with better performance.

## Methods

### Data preprocessing

Following the discussion of [14], we introduce two modifications into the definition Eq. (2) of the frequency counts  $f_i(a)$  and  $f_{ij}(a, b)$ :

- *Pseudocount regularization*: Some amino-acid combinations  $(a, b)$  do not exist in column pairs  $(i, j)$ , even if  $a$  is found in  $i$ , and  $b$  in  $j$ . This would formally lead to infinitely large coupling constants, and the covariance matrix  $C$  becomes non invertible. This divergence can be avoided by introducing a pseudocount  $\tilde{\nu}$ , which adds to the occurrence counts of each amino acid in each column of the MSA.
- *Reweighting*: The sampling of biological sequences is far from being i.i.d., it is biased by the phylogenetic history of the proteins and by the human selection of sequenced species. This bias will introduce global correlations. To reduce this effect, we decrease the statistical weight of sequences having many similar ones in the MSA. More precisely, the weight of each sequence is defined as the inverse number of sequences within Hamming distance  $d_H < xL$ , with an arbitrary but fixed  $x \in (0, 1)$ :

$$w_m = \frac{1}{|\{n | 1 \leq n \leq M; d_H[(a_1^n, \dots, a_L^n), (a_1^m, \dots, a_L^m)] \leq xL\}|} \quad (25)$$

for all  $m = 1, \dots, M$ . The weight equals one for isolated sequences, and becomes smaller the denser the sampling around a sequence is. Note that  $x = 0$  would account to removing double counts from the MSA. The total weight

$$M_{eff} = \sum_{m=1}^M w_m \quad (26)$$

can be interpreted as the effective number of independent sequences.

With these two modifications, frequency counts become

$$f_i(a) = \frac{1}{M_{eff} + \tilde{\nu}} \left[ \frac{\tilde{\nu}}{q} + \sum_{m=1}^M w_m \delta_{a, a_i^m} \right] \quad (27)$$

$$f_{ij}(a, b) = \frac{1}{M_{eff} + \tilde{\nu}} \left[ \frac{\tilde{\nu}}{q^2} + \sum_{m=1}^M w_m \delta_{a, a_i^m} \delta_{b, a_j^m} \right]. \quad (28)$$

Values  $\tilde{\nu} \simeq M_{eff}$  and  $x \simeq 0.2$  were found to work optimally across many protein families [14], we use these values. Besides these modifications, the Hopfield-Potts-model learning is performed as explained before.

### Gauge invariance of Hopfield-Potts model

Amino-acid frequencies are not independent numbers. For instance, on each site  $i$ , the  $q$  amino-acid frequencies add up to one,

$$\sum_{a=1}^q f_i(a) = 1. \quad (29)$$

As a consequence of (29), the Potts model in Eq. (7) has – in physics language – a gauge invariance: any function  $g_i(a)$  can be added to  $e_{ij}(a, b)$  and, simultaneously, be subtracted from  $h_i(a)$  without changing the value of  $P$ . As in [14], we fix the gauge by setting

$$e_{ij}(a, q) = e_{ij}(q, a) = h_i(q) = 0 \quad (30)$$

for all  $i, j$  and all  $a$ . This condition removes completely the gauge freedom, and will be kept throughout the main paper. The parameters to be computed are therefore the couplings  $e_{ij}(a, b)$  and the fields  $h_i(a)$  with  $1 \leq a, b \leq q - 1$ .

An different choice for the gauge is proposed in Supporting Information, and leads to quantitatively equivalent predictions for the pattern structures and the contact map.

## Mean-field theory for determining the Hopfield-Potts patterns

The MaxEnt approach underlying DCA can be rephrased in a Bayesian framework. Assume the model to be given by Eq. (7), and assume the sequences in the MSA to be independently and identically sampled from  $P$ . The probability of the alignment for given model parameters (couplings and fields) is then given by

$$P[A|\{e_{ij}(a, b), h_i(a)\}] = \prod_{m=1}^M P(a_1^m, \dots, a_L^m). \quad (31)$$

Plugging in Eq. (7) and defining the log-likelihood of the model parameters given the MSA  $A$ , we find

$$\begin{aligned} \mathcal{L}[\{e_{ij}(a, b), h_i(a)\}|A] &= \frac{1}{M} \log P[A|\{e_{ij}(a, b), h_i(a)\}] \\ &= \sum_{i < j} \sum_{a, b} e_{ij}(a, b) f_{ij}(a, b) + \sum_{i, a} h_i(a) f_i(a) - \log \mathcal{Z}(\{e_{ij}(a, b), h_i(a)\}) \end{aligned} \quad (32)$$

One can readily see that the parameters  $\{e_{ij}(a, b), h_i(a)\}$  maximizing  $\mathcal{L}$  are solutions of Eqs. (4) and (5). The corresponding value for the maximum of  $\mathcal{L}$  coincides with the opposite of the entropy,  $-S[P]$ , for the MaxEnt distribution given by Eq. (7).

Following the study of the Ising model case ( $q = 2$ ) in [34], mean-field theory can be used to derive an approximate expression for the log-likelihood  $\mathcal{L}$  (32) when the couplings are chosen to obey Hopfield's prescription, Eq. (17). Calculations are presented in the Supporting Information (Sec. I). After optimization over the fields, we are left with the log-likelihood for the patterns only,

$$\mathcal{L}[\{\xi\}|A] = \sum_i \sum_{a=1}^q f_i(a) \log f_i(a) + \frac{1}{2L} \sum_{\mu, ij, ab} \xi_{ia}^\mu C_{ij}(a, b) \xi_{jb}^\mu + \frac{1}{2} \log \left[ 1 - \frac{1}{L} \sum_{i, ab} \xi_{ia}^\mu C_{ii}(a, b) \xi_{ib}^\mu \right] \quad (33)$$

Note that the first term contains a sum over Potts states running up to  $q$  (and not only to  $q - 1$  as in the other expressions), so we find the trivial result that, for  $p = 0$  (no couplings), the likelihood is the negative of the sum of all single-column entropies. The optimal patterns, *i.e.* those optimizing the log-likelihood  $\mathcal{L}$  are given by Eq. (18). The total log-likelihood corresponding to this selection reads:

$$\mathcal{L}(p) = \sum_{i=1}^L \sum_{a=1}^q f_i(a) \log f_i(a) + \sum_{\{\mu | \lambda_\mu \notin (\ell_-, \ell_+)\}} \Delta \mathcal{L}(\lambda_\mu), \quad (34)$$

where function  $\Delta \mathcal{L}$  is defined in Eq. (21), and the bounds  $\ell_-, \ell_+$  are defined in the Results Section.

The solution given in Eq. (18) is defined up to a rotation in the pattern space, *i.e.* up to multiplication of all patterns with an orthogonal  $(p \times p)$ -matrix,  $\mathcal{O}$ . Indeed, the patterns  $\xi_{ia}^\mu$  and their rotated counterparts  $\hat{\xi}_{ia}^\mu = \sum_\nu \mathcal{O}^{\mu\nu} \xi_{ia}^\nu$  define the same set of couplings  $e_{ij}(a, b)$  through Eq. (17). Note that this gauge invariance is specific to the Hopfield model, and should not be mistaken for the gauge invariance of the Potts model discussed in the Results Sections. We eliminate this arbitrariness according to the following procedure, detailed in the Supporting Information. Our selection corresponds to the case where patterns are added one after the other, starting with the best possible single pattern, followed by the second best (orthogonal to the first one when single-site correlations  $C_{ii}(a, b)$  are factored out) etc.

## Error bars on patterns

From a Bayesian point of view, the patterns can fluctuate around the above values according to their posterior distribution. If the prior distribution over the patterns is uniform, we can compute the Fisher information matrix through

$$\mathcal{I}_{\mu ia, \nu jb} = -\frac{\partial^2 \mathcal{L}}{\partial \xi_{ia}^\mu \partial \xi_{jb}^\nu}, \quad (35)$$

where the derivatives are taken around the optimal patterns (18). The deviations of the patterns from their optimal (most likely) values can be estimated from the inverse matrix of  $\mathcal{I}$ . In particular, let us call  $\delta \xi_{ia}^\mu$  the difference between the  $(i, a)$ -component of pattern  $\mu$  and its most likely value given by Eq. (18). If the number of sequences,  $M$ , is large enough,  $\delta \xi_{ia}^\mu$  is distributed as a normal variable, with zero average and variance

$$\langle (\delta \xi_{i,a}^\mu)^2 \rangle = \frac{1}{M} \mathcal{I}_{\mu ia, \mu ia}. \quad (36)$$

The calculation of the variance can be found in Supporting Information, with the result

$$\langle (\delta \xi_{i,a}^\mu)^2 \rangle = \frac{1}{M} \left[ \frac{(\tilde{v}_{ia}^\mu)^2}{2\lambda_\mu(\lambda_\mu - 1)} + \sum_{A \leq p \text{ \& } A \neq \mu} \frac{(\tilde{v}_{ia}^A)^2 (\lambda_\mu - 1) \lambda_A}{(\lambda_\mu |\lambda_A - 1| + \lambda_A |\lambda_\mu - 1|)^2} + \sum_{A > p} \frac{(\tilde{v}_{ia}^A)^2}{\lambda_\mu - \lambda_A} \right]. \quad (37)$$

Components  $\tilde{v}_{ia}^\mu$  have been defined in Eq. (19). Note that the second sum in Eq. (37) runs over all the eigenvectors of  $\Gamma$  with eigenvalues smaller than the ones corresponding to the inferred patterns ( $A > p$ ), while the first sum runs over the top  $p$  eigenvectors (except eigenvector  $\mu$ ).

The above expression is correct when all selected patterns are attractive. Assume now that, say,  $p_- \geq 1$  repulsive patterns (corresponding to the smallest  $p_-$  eigenvalues) and  $p_+$  attractive patterns are retained. Expression (37) is still valid for an attractive pattern, *i.e.* such that  $\mu \leq p_+$ , upon changing condition ( $A \leq p$  &  $A \neq \mu$ ) into ( $A \leq p_+$  &  $A \neq \mu$ , or  $A \geq L(q-1) - p_-$ ) and condition ( $A > p$ ) into ( $p_+ < A < L(q-1) - p_-$ ). For a repulsive pattern, *i.e.* such that  $\mu > L(q-1) - p_-$ , formula (37) holds upon changing condition  $A \leq p$  into ( $A \leq p_+$ , or  $A \geq L(q-1) - p_-$  &  $A \neq \mu$ ) and condition ( $A > p$ ) into ( $p_+ < A < L(q-1) - p_-$ ).

Knowledge of the uncertainties over the patterns allows us to define a Bayesian criterion for pattern selection. Informally speaking, patterns whose components have strong deviations around their most likely values, that is, of the order of the pattern components themselves cannot be considered as reliable and should be discarded. Therefore, for each pattern  $\mu$ , we consider the ratio of the squared fluctuations to the squared norm of the pattern,

$$\rho^\mu = \frac{\sum_{ia} \langle (\delta \xi_{i,a}^\mu)^2 \rangle}{\sum_{ia} (\xi_{i,a}^\mu)^2}. \quad (38)$$

Note that  $\rho^\mu$  is real and positive for both types of patterns (attractive or repulsive). We will decide that the pattern is reliable if the ratio  $\rho^\mu$  is smaller than some arbitrary error threshold, say, 1 or 2/3 [34]. Exemplary results for one protein family (response regulator domain) are given in the supplementary Fig. 13. Note that the error bars depend on the error threshold itself, smaller error thresholds lead to increased errors of the selected patterns. As a consequence, pattern selection according to the uncertainty of patterns is a self-consistent criterion, which can be solved in an iterative way.

As can be seen in the inset in supp. Fig. 13, log-likelihood and error are in an almost one-to-one relation, deviations appear only for the first few patterns. Therefore both selection criteria lead, when the arbitrary thresholds are chosen coherently, to almost equivalent results, and we will concentrate on the simpler to handle maximum-likelihood criterium in the remainder of this article.

## Contact prediction from couplings

Intuitively, residue position pairs with strong direct couplings are our best predictions for native contacts in the protein structure. To measure 'coupling strength', we need, however, to map the inferred coupling matrices  $e_{ij}$  onto a scalar parameter, for each  $1 \leq i < j \leq L$ . Whereas previous works on DCA have mainly used the so-called direct information [12,14], it was recently observed that a different score actually improves the contact prediction starting from the same model parameters  $\{e_{ij}(a, b)\}$  [41]. To this end, we introduce the Frobenius norm

$$F_{ij} = \|e'_{ij}\|_2 = \sqrt{\sum_{a,b=1}^q \tilde{e}_{ij}(a, b)^2} \quad (39)$$

of the linearly transformed coupling matrices

$$\tilde{e}_{ij}(a, b) = e_{ij}(a, b) - e_{ij}(\cdot, b) - e_{ij}(a, \cdot) + e_{ij}(\cdot, \cdot) , \quad (40)$$

where ' $\cdot$ ' denotes average over all amino acids and the gap in the concerned position. According to the above discussion, this corresponds to another gauge of the Hopfield-Potts model, more precisely to the gauge minimizing the Frobenius norm of each coupling matrix [12]. Further more, the norm is adjusted by an *average product correction* (APC) term, introduced in [11] to suppress effects from phylogenetic bias and insufficient sampling. Incorporating also this correction, we get our final scalar score:

$$F_{ij}^{APC} = F_{ij} - \frac{F_{\cdot j} F_{i \cdot}}{F_{\cdot \cdot}} , \quad (41)$$

where the ' $\cdot$ ' now indicates a position average.

Sorting column pairs  $(i, j)$  by decreasing values of  $F_{ij}^{APC}$  calculated using standard mean-field DCA was shown to give accurate predictions for residue contacts in various proteins, *i.e.* in the case where all possible patterns are included ( $p = L(q-1)$ ) in Eq. (17). The Results Section shows how the performance in contact prediction varies when the number of patterns is  $p \ll L(q-1)$ .

## Acknowledgments

We are grateful to R. Ranganathan and O. Rivoire for interesting discussions. This work was facilitated by the Initiative for the Theoretical Sciences at The Graduate Center of the City University of New York. S.C. and R.M. acknowledge the hospitality of the Simons Center for Systems Biology, Institute for Advanced Study, Princeton, where an initial part of this work was done. This work benefited from the European Community FP7/2007-2013/grant agreement n290038 (Netadis).

## References

1. Pagani I, Liolios K, Jansson J, Chen I, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 40: D571.
2. The Uniprot Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40: D71.
3. Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate JG, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40: D290.

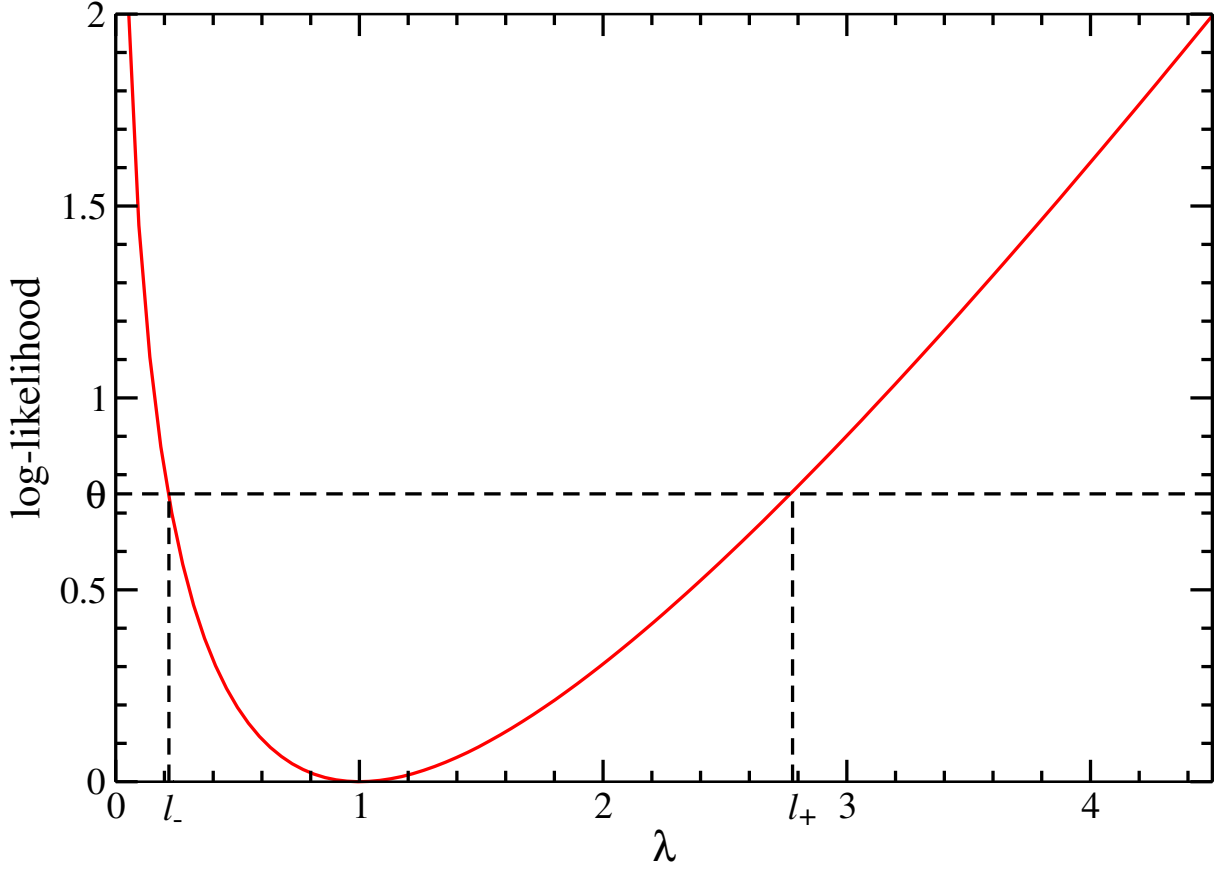


4. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The protein data bank at 40: Reflecting on the past to prepare for the future. *Structure* 20: 391 - 396.
5. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins: Struct, Funct, Genet* 18: 309.
6. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286: 295-299.
7. Lapedes AS, Giraud BG, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: Phylogenetic and structural effects. *Lecture Notes-Monograph Series: Statistics in Molecular Biology and Genetics* 33: pp. 236-256.
8. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics* 56: 211-221.
9. Socolich M, Lockless SW, Lee HL, Gardner K, Ranganathan R (2005) Evolutionary Information for Specifying a Protein Fold. *Nature* 437: 512-518.
10. Russ W, Lowery D, Mishra P, Yaffe M, Ranganathan R (2005) Natural-like Function in Artificial WW Domains. *Nature* 437: 579-583.
11. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24: 333.
12. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 106: 67.
13. Burger L, van Nimwegen E (2010) Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLoS Comput Biol* 6: E1000633.
14. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108: E1293.
15. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. *Proteins: Struct, Funct, Bioinf* 79: 1061.
16. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28: 184.
17. Jaynes ET (1957) Information Theory and Statistical Mechanics. *Physical Review Series II* 106: 620630.
18. Jaynes ET (1957) Information Theory and Statistical Mechanics II. *Physical Review Series II* 108: 171190.
19. Schneidman E, Berry M, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440: 1007-1012.
20. Cocco S, Leibler S, Monasson R (2009) Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc Natl Acad Sci U S A* 106: 14058-62.

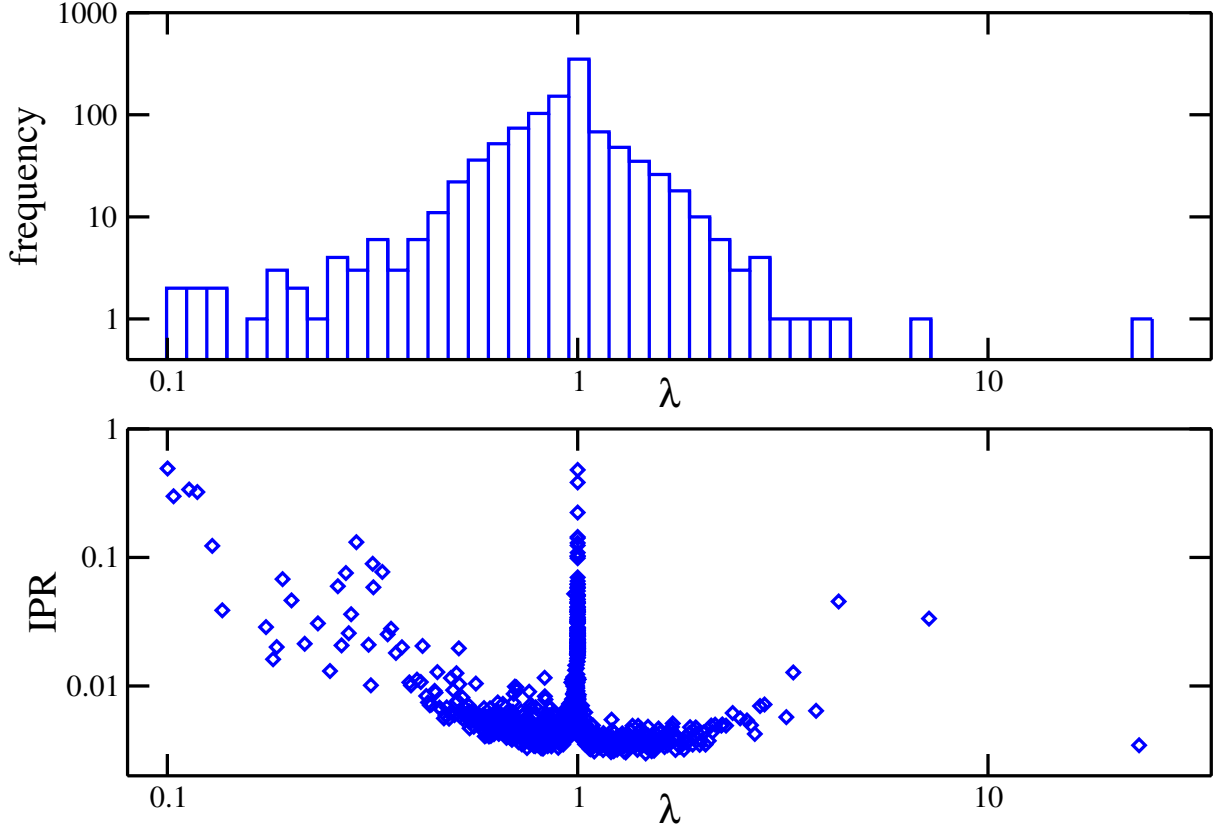
21. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Nat Acad Sci* 103: 19033-19038.
22. Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, et al. (2012) Statistical mechanics for natural flocks of birds. *Proc Nat Acad Sci* .
23. Marks DS, Colwell LJ, Sheridan RP, Hopf TA, Pagnani A, et al. (2011) 3D Protein Structure Predicted from Sequence. *arXiv:11105091* .
24. Sulkowska JJ, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci* 109: 10340-10345.
25. Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences* 109: E1540-E1547.
26. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149: 1607-1621.
27. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA* 106: 22124.
28. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci USA* 109: 10148.
29. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2: 559-572.
30. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* 138: 774-786.
31. Reynolds KA, McLaughlin RN, R R (2011) Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* 147: 1564-1575.
32. Bai Z, Silverstein JW (2009) Spectral analysis of large dimensional random matrices. London: Springer.
33. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79: 2554-2558.
34. Cocco S, Monasson R, Sessak V (2011) High-Dimensional Inference with the generalized Hopfield Model: Principal Component Analysis and Corrections. *Physical Review E* 83: 051123.
35. Wu FY (1982) The Potts Model. *Rev Mod Phys* 54: 235-268.
36. Cox T, Cox M (1994) Multidimensional Scaling. London: Chapman & Hall.
37. Nokura K (1998) Spin glass states of the anti-hopfield model. *J Phys A* 31: 7447.
38. Wlodawer A, Walter J, Huber R, Sjolín L (1984) Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and x-ray refinement of crystal form ii. *Journal of Molecular Biology* 180: 301 - 329.

39. Bent CJ, Isaacs NW, Mitchell TJ, Riboldi-Tunncliffe A (2004) Crystal Structure of the Response Regulator 02 Receiver Domain, the Essential YycF Two-Component System of *Streptococcus pneumoniae* in both Complexed and Native States. *J Bacteriol* 186: 2872-2879.
40. Pai EF, Krengel U, Petsko GA, Goody RS, Kabsch W, et al. (1990) Refined crystal structure of the triphosphate conformation of h-Ras p21 at 1.35 Å resolution: implications for the mechanism of gtp hydrolysis. *EMBO J* 9: 2351-2359.
41. Ekeberg M, Lökvist C, Lan Y, Weigt M, E A (2012) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models .

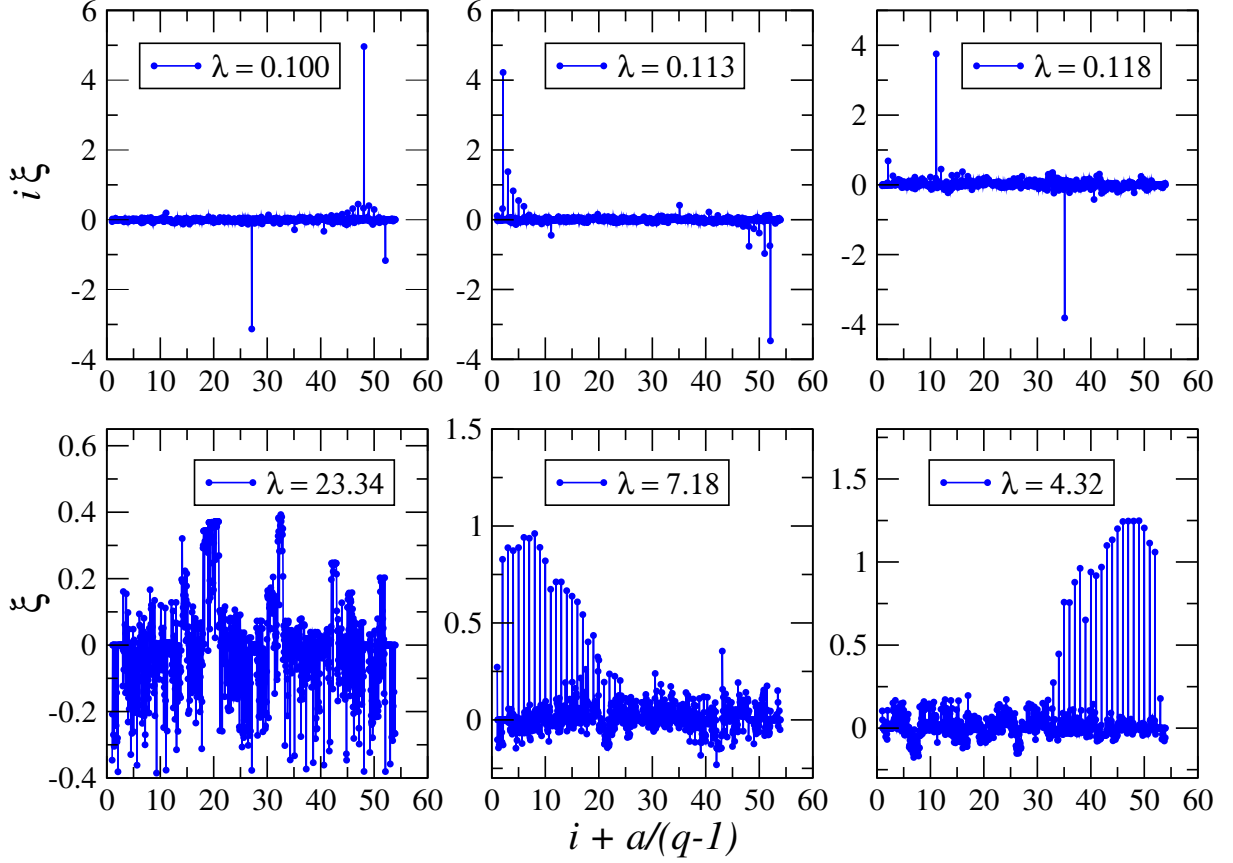
## Figure Legends



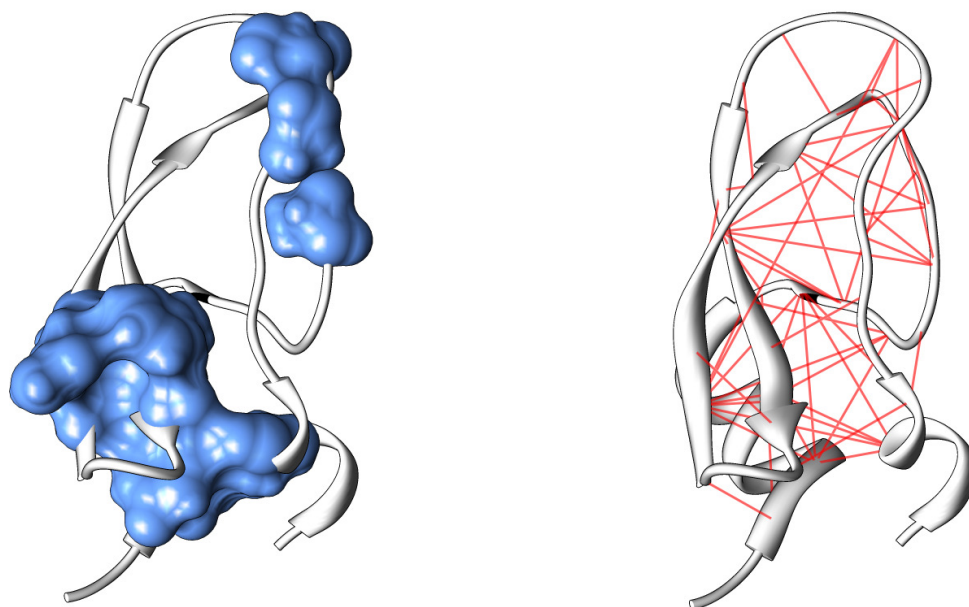
**Figure 1. Pattern selection by maximum likelihood:** Contribution of patterns to the log-likelihood (full red line) as a function of the corresponding eigenvalues  $\lambda$  of the Pearson correlation matrix  $\Gamma$ . To select  $p$  patterns, a log-likelihood threshold  $\theta$  (dashed black line) has to be chosen such that there are exactly  $p$  patterns with  $\Delta\mathcal{L}(\lambda_\mu) > \theta$ . This corresponds to eigenvalues in the left and right tail of the spectrum of  $\Gamma$ .



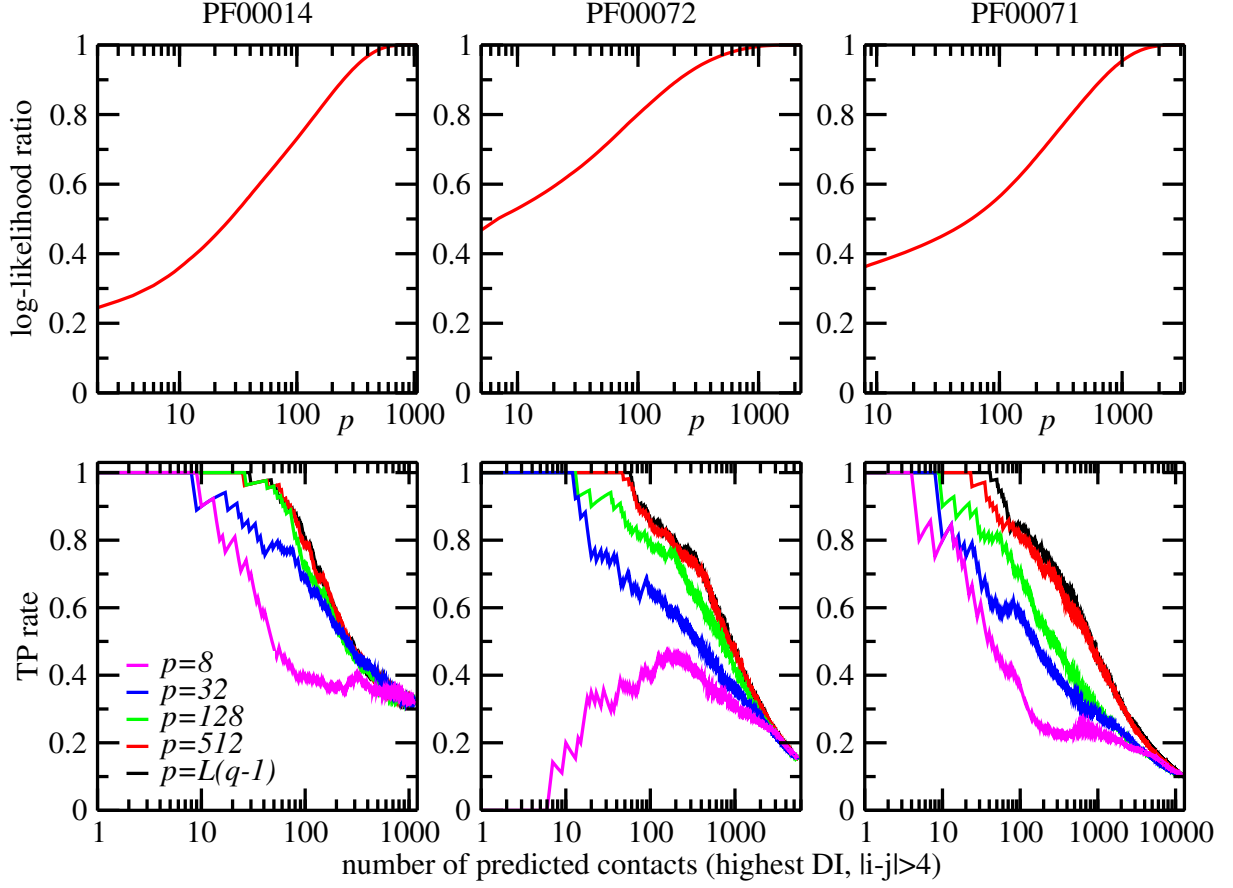
**Figure 2. Eigenvalues and localization for PF00014:** (*upper panel*) The spectral density as a function of the eigenvalues  $\lambda$ , note the existence of few very large eigenvalues, and a pronounced peak in  $\lambda = 1$ . (*lower panel*) The inverse participation ratio of the Hopfield patterns as a function of the corresponding eigenvalue  $\lambda$ . Large IPR characterizes the concentration of a pattern to few positions and amino acids.



**Figure 3. Attractive and repulsive patterns for PF00014:** (*upper panels*) The most localized repulsive patterns (corresponding to the first, third and fourth smallest eigenvalues and inverse participation ratios 0.49, 0.34, 0.32 respectively) are strongly concentrated in pairs of positions. (*lower panels*) The most attractive patterns (corresponding to the three largest eigenvalues); the top pattern is extended, with inverse participation ratio 0.003, while the second and third patterns, with inverse participation ratios 0.033, 0.045 respectively, have essentially non-zero components over the gap symbols only which accumulate on the edges of the sequence. Note the  $x$ -coordinates  $i + a/(q-1)$ ; its integer part is the site index,  $i$ , and the fractional part multiplied by  $q-1$  is the residue value,  $a$ .

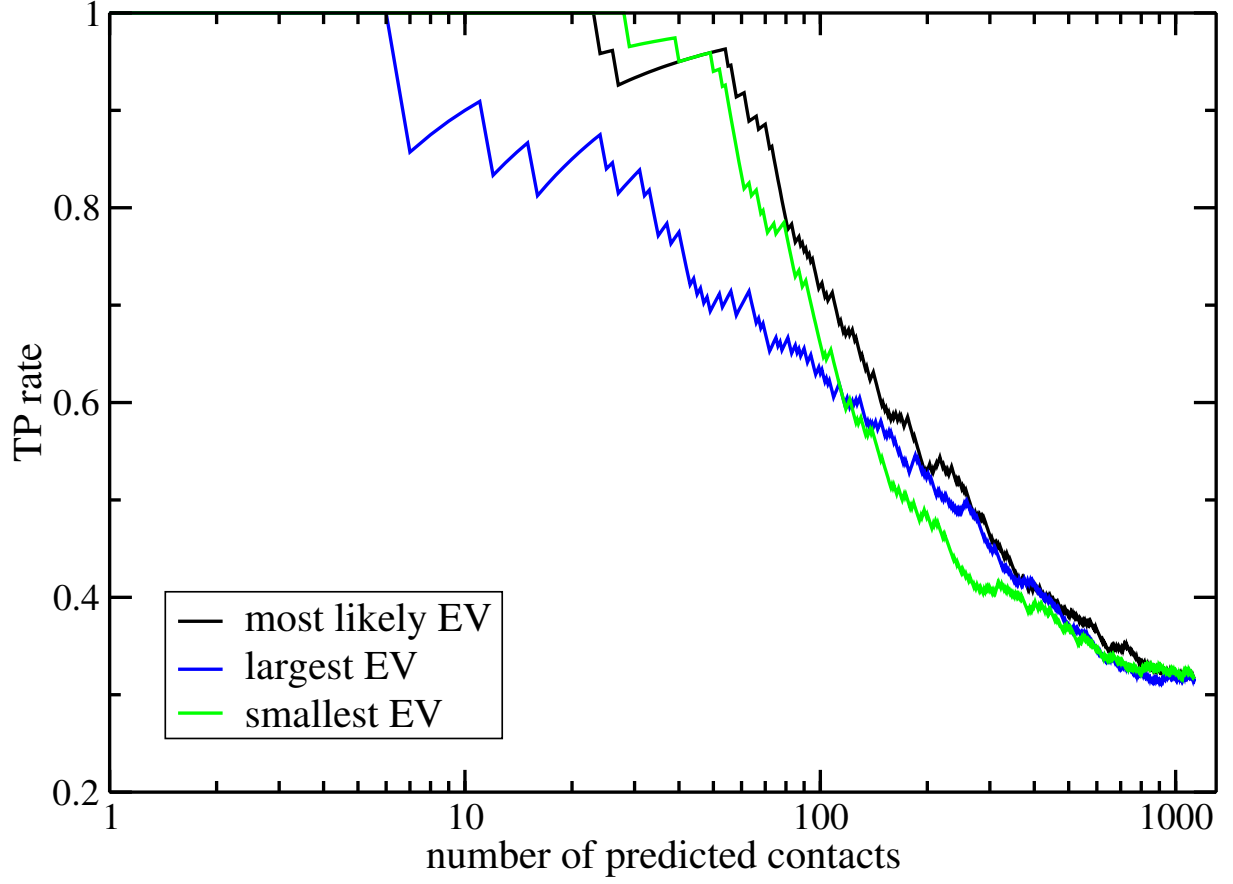


**Figure 4. The principal component and predicted contacts visualized on the 3D structure of the trypsin inhibitor protein domain PF00014.** (A) The 15 positions of largest entries in the most attractive Hopfield pattern (largest eigenvalue of  $\Gamma$ , corresponding to the principal component) are shown in blue, they correspond also to the most conserved sites. Note that, while they are distant along the protein backbone, they cluster into spatially connected components in the folded protein. (B) The 50 residue pairs with strongest couplings (ranked according to the Frobenius norms Eq. 41), with at least 5 positions separation along the backbone, are connected by red lines. Note that they include many pairs between not conserved positions. Only two out of these pairs are not in contact.



**Figure 5. Contact predictions for the three considered protein families.** The upper panels show the fraction of the interaction-based contribution to the log-likelihood as a function of the number  $p$  of selected patterns, it reaches one for  $p = (q - 1)L$ . The lower panels show the TP rates as a function of the predicted residue contacts, for various numbers  $p$  of selected patterns, where selection was done using the maximum-likelihood criterium.





**Figure 6. Contact predictions by attractive and repulsive patterns for PF00014.** TP rates for the contact prediction using purely repulsive resp. attractive patterns, resulting from selecting the 100 smallest [green] resp. largest [blue] eigenvalues. The results are compared to the TP rates obtained by selecting the 100 most likely Hopfield patterns (black).